09/669,680

1. (Currently Amended) A method of clustering documents in datasets comprising:
clustering first documents ~~and~~ in a first dataset to produce first document classes;
creating centroid seeds based on said first document classes; and
clustering second documents in a second dataset using said centroid seeds.

2. (Original) The method in claim 1, wherein said first dataset and said second dataset are related.

3. (Original) The method in claim 1, wherein said clustering of said first documents in said first dataset comprises:
forming a first dictionary of most common words in said first dataset;
generating a first vector space model by counting, for each word in said first dictionary, a number of said first documents in which said word occurs; and
clustering said first documents in said first dataset based on said first vector space model.

4. (Original)The method in claim 3, further comprising generating a second vector space model by counting, for each word in said first dictionary, a number of said second documents in which said word occurs.

5. (Original)The method in claim 4, wherein said creating of said centroid seeds comprises:
classifying said second vector space model using said first document classes to produce a classified second vector space model; and
determining a mean of vectors in each class in said classified second vector space model, wherein said mean comprises said centroid seeds.

6. (Original)The method in claim 4, further comprising:
forming a second dictionary of most common words in said second dataset;
generating a third vector space model by counting, for each word in said second

dictionary, a number of said second documents in which said word occurs; and

clustering said documents in said second dataset based on said third vector space model to produce a second dataset cluster.

7. (Original)The method in claim 6, wherein said clustering of said second documents in said second dataset using said centroid seeds produces an adapted dataset cluster and said method further comprises:

comparing classes in said adapted dataset cluster to classes in said second dataset cluster; and

adding classes to said adapted dataset cluster based on said comparing.

8. (Original)A system for clustering documents in datasets comprising:

a storage having a first dataset and a second dataset;

a cluster generator operative to cluster first documents in said first dataset and produce first document classes; and

a centroid seed generator operative to generate centroid seeds based on said first document classes,

wherein said cluster generator clusters second documents in said second dataset using said centroid seeds.

9. (Original)The system in claim 8, wherein said first dataset and said second dataset are related.

10. (Original)The system in claim 8, further comprising:

a dictionary generator adapted to generate a first dictionary of most common words in said first dataset; and

a vector space model generator adapted to generate a first vector space model by counting, for each word in said first dictionary, a number of said first documents in which said

word occurs,

wherein said cluster generator clusters said documents in said first dataset based on said first vector space model.

11. (Original)The system in claim 10, wherein said vector space model generator generates a second vector space model by counting, for each word in said first dictionary, a number of said second documents in which said word occurs.

12. (Original)The system in claim 11, further comprising a classifier adapted to classify said second documents in said second vector space model using said first document classes to produce a classified second vector space model and adapted to determine a mean of vectors in each class in said classified second vector space model, wherein said mean comprises said centroid seeds.

13. (Original)The system in claim 11, wherein:

said dictionary generator is adapted to generate a second dictionary of most common words in said second dataset,

said vector space model generator is adapted to generate a third vector space model by counting, for each word in said second dictionary, a number of said second documents in which said word occurs, and

said cluster generator is adapted to cluster said second documents in said second dataset based on said third vector space model to produce a second dataset cluster.

14. (Original)The system in claim 13, wherein said cluster generator is adapted to produce an adapted dataset cluster by clustering said second documents in said second dataset using said centroid seeds and said system further comprises:

a comparator adapted to compare classes in said adapted dataset cluster to classes in said second dataset cluster and add classes to said adapted dataset cluster based on said comparing.

15. (Original)A method of clustering documents in a first dataset having first documents and a related second dataset having second documents, said method comprising:

clustering said first documents to produce first document classes;

generating a vector space model of said second documents;

classifying said vector space model of said second documents using said first document classes to produce a classified vector space model; and

determining a mean of vectors in each class in said classified vector space model to produce centroid seeds; and

clustering said second documents using said centroid seeds.

16. (Original)The method in claim 15, wherein said vector space model comprises a second vector space model and said clustering of said first documents in said first data comprises:

forming a first dictionary of most common words in said first dataset; and

generating a first vector space model by counting, for each word in said first dictionary, a number of said first documents in which said word occurs,

wherein said clustering of said first documents in said first dataset is based on said first vector space model.

17. (Original)The method in claim 16, wherein said generating of said second vector space model comprises counting, for each word in said first dictionary, a number of said second documents in which said word occurs.

18. (Original)The method in claim 17, further comprising:

forming a second dictionary of most common words in said second dataset;

generating a third vector space model by counting, for each word in said second dictionary, a number of said second documents in which said word occurs; and

clustering said documents in said second dataset based on said third vector space model

to produce a second dataset cluster.

19. (Original)The method in claim 18, wherein said clustering of said second documents in said second dataset using said centroid seeds produces an adapted dataset cluster and said method further comprises:

comparing classes in said adapted dataset cluster to classes in said second dataset cluster; and

adding classes to said adapted dataset cluster based on said comparing.

20. (Original)A method of clustering documents in related datasets comprising:

forming a first dictionary of most common words in a first dataset;

generating a first vector space model by counting, for each word in said first dictionary, a number of said first documents in which said word occurs; and

clustering said first documents in said first dataset based on said first vector space model to produce first document classes;

generating a second vector space model by counting, for each word in said first dictionary, a number of said second documents in which said word occurs;

classifying said second documents in said second vector space model using said first document classes to produce a classified second vector space model;

determining a mean of vectors in each class in said classified second vector space model to produce centroid seeds; and

clustering second documents in a second dataset using said centroid seeds

21. (Original)The method in claim 20, further comprising:

forming a second dictionary of most common words in said second dataset;

generating a third vector space model by counting, for each word, in said second dictionary, a number of said second documents in which said word occurs; and

clustering said documents in said second dataset based on said third vector space model to produce a second dataset cluster.

22. (Original)The method in claim 21, wherein said clustering of said second documents in said second dataset using said centroid seeds produces an adapted dataset cluster and said method further comprises:

comparing classes in said adapted dataset cluster to classes in said second dataset cluster; and

adding classes to said adapted dataset cluster based on said comparing.

23. (Currently Amended)A program device readable by ~~machines~~ machine tangibly embodying a program of instructions executable by the machine to perform ~~said~~ a method of clustering documents in datasets comprising:

clustering first documents ~~and~~ in a first dataset to produce first document classes;

creating centroid seeds based on said first document classes; and

clustering second documents in a second dataset using said centroid seeds.

24. (Currently Amended )A program device readable by ~~machines~~ machine ,tangibly embodying a program of instructions executable by the machine to perform said method in claim 23, wherein said first dataset and second dataset are related.

25. (Currently Amended)A program device readable by ~~machines~~ machine, tangibly embodying a program of instructions executable by the machine to perform said method in claim 23, wherein said clustering of said first documents in said first dataset comprises:

forming a first dictionary of most common words in said first dataset;

generating a first vector space model by counting, for each word in said first dictionary, a number of said first documents in which said word occurs; and

clustering said first documents in said first dataset based on said first vector space model.

26. (Currently Amended) A program device readable by ~~machines~~ machine , tangibly embodying a program of instructions executable by the machine to perform said method in claim 25, said method further comprising generating a second vector space model by counting, for each word in said first dictionary, a number of said second documents in which said word occurs.

27. (Currently Amended) A program device readable by ~~machines~~ machine , tangibly embodying a program of instructions executable by the machine to perform said method in claim 26, wherein said creating of said centroid seeds comprises:

classifying said second vector space model using said first document classes to produce a classified second vector space model; and

determining a mean of vectors in each class in said classified second vector space model, wherein said mean comprises said centroid seeds.

28. (Currently Amended) A program device readable by ~~machines~~ machine, tangibly embodying a program of instructions executable by the machine to perform said method in claim 26, said method further comprising:

forming a second dictionary of most common words in said second dataset;

generating a third vector space model by counting, for each word in said second dictionary, a number of said second documents in which said word occurs; and

clustering said documents in said second dataset based on said third vector space model to produce a second dataset cluster.

29. (Currently Amended) A program device readable by ~~machines~~ machine, tangibly embodying a program of instructions executable by the machine to perform said method in claim 28, wherein said clustering of said second documents in said second dataset using said centroid seeds produces an adapted dataset cluster and said method further comprises:

comparing classes in said adapted dataset cluster to classes in said second dataset cluster;

09/669,680

and

adding classes to said adapted dataset cluster based on said comparing.

30. (Canceled).